Abstract

Correction of the articles in English texts is approached as an article generation task, i.e. each noun phrase is assigned with a class corresponding to the definite, indefinite or zero article. Supervised machine learning methods are used to first replicate and then improve upon the best reported result in the literature known to the author. By feature engineering and a different choice of the learning method, about 34% drop in error is achieved. The resulting model is further compared to the performance of expert annotators. Although the comparison is not straightforward due to the differences in the data, the results indicate the performance of the trained model is comparable to the human-level performance when measured on the in-domain data. On the other hand, the model does not generalize well to different types of data. Using a large-scale language model to predict an article (or no article) for each word of the text has not proved successful.